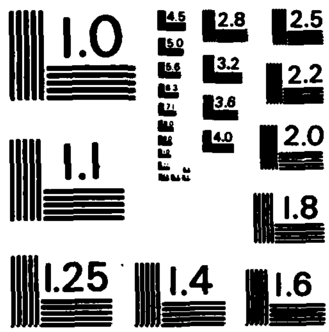
AD-A162 169 AN INEQUALITY CONCERNING THE DEVIATION BETWEEN THEORETICAL AND EMPIRICAL..(U) PITTSBURGH UNIV PA CENTER FOR MULTIVARIATE ANALYSIS L ZHAO AUG 85 1/1

UNCLASSIFIED TR-85-30 AFOSR-TR-85-0978 F49620-85-C-0008 F/G 12/1 NL

END

FILMED

DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AN INEQUALITY CONCERNING THE DEVIATION BETWEEN THEORETICAL AND EMPIRICAL DISTRIBUTIONS*

Lin-Cheng Zhao
Center for Multivariate Analysis

Center for Multivariate Analysis

University of Pittsburgh

AN INEQUALITY CONCERNING THE DEVIATION BETWEEN THEORETICAL AND EMPIRICAL DISTRIBUTIONS*

Lin-Cheng Zhao
Center for Multivariate Analysis

August 1985

Technical Report No. 85-30

Center for Multivariate Analysis
University of Pittsburgh
515 Thackeray Hall
Pittsburgh, PA 15260

*Part of the work of this author is sponsored by the Air Force Office of Scientific Research (AFSC) under contract F49620-85-C-0008. The United States Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation herein.

1. INTRODUCTION

The result. Let $x_1,\ldots,x_r$ be r points in R^d, and $\mathcal{A}$ be a class of Borel sets in R^d. Denote by $\Delta^{\mathcal{A}}(x_1,\ldots,x_r)$ the number of distinct sets in $\{\{x_1,\ldots,x_r\} \cap A,\ A \in \mathcal{A}\}$. Define

$$m^{\mathcal{A}}(r) = \max_{x_1,\ldots,x_r \in R^d} \Delta^{\mathcal{A}}(x_1,\ldots,x_r).$$

Vapnik and Chervonenkis (1971) showed that either $m^{\mathcal{A}}(r) = 2^r$ for any positive integer r or $m^{\mathcal{A}}(r) \leq r^s+1$, where s is the smallest k such that $m^{\mathcal{A}}(k) \neq 2^k$. A class of sets $\mathcal{A}$ for which the latter case holds will be called a V-C class with index s.

Suppose that μ is a probability measure on R^d. Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random vectors with common distribution μ, and μ_n be the empirical distribution of $X_1,\ldots,X_n$. Denote a "distance" between μ_n and μ by

$$D_n(\mathcal{A},\mu) = \sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|.$$

Throughout this paper we assume that $D_n(\mathcal{A},\mu)$, $\sup_{A\in\mathcal{A}}|\mu_n(A)-\mu_{2n}(A)|$ and $\sup_{A\in\mathcal{A}} \mu_n(A)$ are all random variables. We shall prove the following

<u>Theorem 1</u>. Let $\mathcal{A}$ be a V-C class with index s such that

$$\sup_{A\in\mathcal{A}} \mu(A) \leq \delta \leq 1/8. \tag{1}$$

Then for any $\varepsilon > 0$ we have

$$P\{D_n(\mathcal{A},\mu)>\varepsilon\} \leq 5(2n)^s \exp(-n\varepsilon^2/(91\delta+4\varepsilon)) \tag{2}$$

$$+ 7(2n)^s \exp(-\delta n/68)$$

$$+ 2^{2+s} n^{1+2s} \exp(-\delta n/8),$$

provided $n \geq \max\left(12\sigma/\varepsilon^2,\ 68(1+s)(\log 2)/\delta\right)$.

The proof of (2) is based on an important inequality proved by Devroye and Wagner (1980).

2. HISTORICAL NOTES

A few remarks concerning this inequality are in order. In 1971, Vapnik and Chervonenkis proved that, for any $\varepsilon > 0$

$$P\{D_n(A,\mu) > \varepsilon\} \leq 4\exp(-n\varepsilon^2/8)\ E\Delta^A(X_1,\ldots,X_{2n}). \tag{3}$$

This inequality is quite general since no restrictions such as (1) are imposed. In using this inequality, an extimate of $m^A(n)$ must be given, see, for example, Gaenssler and Stute (1979), Wenocur and Dudley (1981).

The weakness of (3) lies in the fact that, in many applications $\varepsilon = \varepsilon_n \to 0$ as $n\to\infty$. In this case $n\varepsilon_n^2$ may not tend to ∞ or tend to ∞ very slowly. For this reason, the inequality proved by Devroye and Wagner (1980) is sometimes more useful. They proved that, if $\sup_A \mu(A) \leq \delta \leq \frac{1}{4}$, then for any $\varepsilon > 0$

$$\begin{aligned} P\{D_n(A,\mu) > \varepsilon\} &\leq 4m^A(2n)\exp(-n\varepsilon^2/(64\delta+4\varepsilon)) \\ &+ 2P\{\operatorname{Sup}_A \mu_{2n}(A) > 2\delta\} \end{aligned} \tag{4}$$

for $n \geq 8\delta/\varepsilon^2$. If we further have

$$\operatorname*{Sup}_{A\varepsilon A}\ \operatorname*{Sup}_{x,y\varepsilon A} ||x-y|| \leq \rho < \infty$$

and

$$\operatorname*{Sup}_{x\varepsilon R^d} \mu(S(x,\rho)) \leq \delta \leq \tfrac{1}{4}, \tag{5}$$

here $||\cdot||$ is the L_2 or L_∞ norm in R^d, and $S(x,\rho)$ is the closed ball with radius ρ

centered at x, then

$$P\{D_n(A,\mu)>\varepsilon\} \leq 4m^A(2n)\exp(-n\varepsilon^2/(64\delta+4\varepsilon)) + 4n\exp(-n\delta/10) \tag{6}$$

for $n \geq \max(1/\delta, 8\delta/\varepsilon^2)$.

This inequality is most useful when A is the class of balls with the same diameter (norm L_2 or L_∞). Otherwise δ may be much larger than $\mathrm{Sup}_A\, \mu(A)$, and (6) gives no improvement over (3). Chen and Zhao (1984) made an essential improvement in the one-dimensional case:

Let A be a class of intervals in R^1, satisfying $\sup_{I\in A} \mu(I) \leq \delta \leq 1$.

Then there exists positive absolute constants $C_0, C_1, \ldots, C_4$ such that for any $\varepsilon > 0$

$$P\{\sup_{I\in A}|\mu_n(I) - \mu(I)| > \varepsilon\} \leq C_1\varepsilon^{-1}\sqrt{\delta/n}\exp(-C_2 n\varepsilon^2/\delta) + C_3\exp(-C_4 n\varepsilon), \tag{7}$$

provided $n/\log n > C_0/\varepsilon$.

The proof of (7) relies on a result concerning the strong approximation to Brownian bridge of the empirical process on R^1. The argument fails in the general case $d > 1$. The inequality (2), to be proved in the next section, gives a satisfactory generalization to the case $d \geq 1$.

3. PROOF OF THEOREM 1

Set

$$\delta_j = 2^{2^{-1}+2^{-2}+\dots+2^{-j}}\delta, \quad j=1,2,\dots,r,$$

where r will be chosen later. Then

$$\delta < \delta_1 < \delta_2 < \dots < \delta_r < 2\delta \leq \tfrac{1}{4}$$

When $n \geq 12\delta/\varepsilon^2$ we have $n \geq 8\delta_1/\varepsilon^2$. From (4), the definition of V-C class and the fact that

$$\mathrm{Sup}_A \ \mu(A) \leq \delta_1 \leq \tfrac{1}{4},$$

it follows that

$$P\{D_n(A,\mu)>\varepsilon\} \leq 4\{(2n)^s+1\} \exp(-n\varepsilon^2/(64\delta_1+4\varepsilon))$$

$$+ 2P\{\mathrm{Sup}_A \ \mu_{2n}(A) > 2\delta_1\}$$

$$\leq 5(2n)^s\exp(-n\varepsilon^2/(64\sqrt{2}\delta+4\varepsilon)) + 2P\{D_{2n}(A,\mu) > \delta_1\},$$

provided $n \geq 12\delta/\varepsilon^2$.

When $\delta n \geq 68(1+s)\log 2$, we have $2^{j-1}n \geq 8\delta_j/\delta_{j-1}^2$ for $j = 2,3,\dots r$. As before, from (4) and $\mathrm{Sup}_A \ \mu(A) \leq \delta_2 \leq \frac{1}{4}$, it follows that

$$P\{D_n(A,\mu) > \varepsilon\} \leq 5(2n)^s\exp(-n\varepsilon^2/(91\delta+4\varepsilon))$$

$$+(2\cdot 5)(2\cdot 2n)^s\exp(-2n\delta_1^2/(64\delta_2+4\delta_1))$$

$$+ \ 2^2 \ P\{D_{2^2 n}(A,\mu) > \delta_2\},$$

provided $n \geq \max(68(1+s)\log 2/\delta,\ 12\delta/\varepsilon^2)$.

Using (4) and $\sup_A \mu(A) \leq \delta_j \leq \frac{1}{4}$ repeatedly, we obtain

$$P\{D_n(A,\mu) > \varepsilon\} \leq 5(2n)^s \exp(-n\varepsilon^2/(91\delta+4\varepsilon))$$
$$+ \sum_{j=1}^{r-1} 2^j \cdot 5(2^j \cdot 2n)^s \exp(-2^j n\delta_j^2/(68\delta_{j+1})) \tag{8}$$
$$+ 2^r P\{D_{2^r n}(A,\mu) > \delta_r\} \triangleq J_{1,n} + J_{2,n} + J_{3,n},$$

provided $n \geq \max(68(1+s)\log 2/\delta,\ 12\delta/\varepsilon^2)$.

It is easy to see that

$$2^j \delta_j^2/\delta_{j+1} \geq 2j\delta, \quad j=1,\ldots, r-1. \tag{9}$$

Hence it follows from (8), (9) and $2^{1+s} \leq e^{\delta n/68}$ that

$$J_{2,n} \leq 5(2n)^s \sum_{j=1}^{r-1} 2^{(1+s)j} \cdot \exp(-2^j n\delta_j^2/(68\delta_{j+1}))$$
$$\leq 5(2n)^s \sum_{j=1}^{\infty} (2^{1+s})^j \exp(-2j\delta n/68)$$
$$\leq 5(2n)^s \sum_{j=1}^{\infty} \exp(-j\delta n/68)$$
$$= 5(2n)^s e^{-\delta n/68}(1-e^{-\delta n/68})^{-1}$$
$$\leq 5(2n)^s(1-2^{-(1+s)})^{-1} e^{-\delta n/68}$$
$$\leq 7(2n)^s \exp(-\delta n/68),$$

where $s \geq 1$ is invoked.

When $\delta n \geq 68(1+s)\log 2$, we have $2^r n\delta_r \geq 2$. By (3)

$$J_{3,n} \leq 2^{r+1}((2^{r+1}n)^s+1)\exp(-2^r n\delta_r^2/8). \tag{11}$$

Take $r = r_n$ to be an integer such that $n/2 < 2^r \leq n$. When $\delta n \geq 68(1+s)\log 2$, we have $n^2\delta_r^2 \geq 2$, $n\delta_r \geq \sqrt{2}$ and $n\delta_r^2 \geq 2\delta$. By (11) we have

$$\begin{aligned} J_{3,n} &\leq 2n((2n^2)^s+1)\exp(-n^2\delta_r^2/16) \\ &\leq 4n(2n^2)^s\exp(-\delta n/8). \end{aligned} \tag{12}$$

Formula (2) follows from (8), (10) and (12). The theorem is proved.

4. APPLICATIONS

Theorem 1 has some applications in strong convergence problems involving the uniform deviation between frequencies and probabilities of a class of events. As an example, we consider the nearest neighbor (NN) density estimates proposed by Loftsgarden and Quesenberry (1965). Suppose that X is a R^d-valued random vectors with distribution μ and unknown density function f. The so called NN estimate of f(x) has the form

$$\hat{f}_n(x) = {}^k/\{n(2a_n(x))^d\},\ x = (x^{(1)},\ldots,x^{(d)})\varepsilon R^d, \tag{13}$$

where $k = k_n \leq n$ is a positive integer chosen in advance, $a_n(x)$ is the smallest $a > 0$ such that the cube $[x-a,x+a] = \Pi_{i=1}^d [x^{(i)}-a,x^{(i)}+a]$ contains at least k sample points. As an application of Theorem 1, we prove a theorem about the convergence rate of $\mathrm{Sup}_{x\varepsilon R^d} |\hat{f}_n(x) - \hat{f}(x)|$

In the sequel, we use c, α, c_1, c_2, ... for some positive constants independent of n and x. For $x = (x^{(1)},\ldots,x^{(d)})\varepsilon R^d$, $y = (y^{(1)},\ldots,y^{(d)})\varepsilon R^d$, write $f'(x)(y-x) = \sum_{i=1}^d \frac{\partial f}{\partial x^{(i)}} (y^{(i)}-x^{(i)})$, and take $||y - x|| = \max_{1\leq i\leq d}|y^{(i)} - x^{(i)}|$. We say that the density function f belongs to λ-class for some $\lambda\varepsilon(0,2]$, if $\lambda\varepsilon(0,1]$ and $|f(y) - f(x)| \leq C||y-x||^\lambda$ for any $x,y\varepsilon R^d$, or $\lambda\varepsilon(1,2]$ and, f are bounded and

$$|f(y) - f(x) - f'(x)(y-x)| \leq C||y - x||^\lambda$$

for any $x,y\varepsilon R^d$. We have

<u>Theorem 2</u>. Suppose that f belongs to λ-class for some $\lambda\varepsilon(0,2]$. Take $k = o(n)$ and

$$^k/n \geq \beta \left(\frac{\log n}{n}\right)^{(d+\lambda)/(d+3\lambda)} \tag{14}$$

where $\beta > 0$ is any given constant. Then

$$\lim_{n\to\infty} \sup\{(^n/k)^{\lambda/(d+\lambda)} \operatorname{Sup}_x |\hat{f}_n(x) - \dot{f}(x)|\} \leq C \text{ a.s.} \tag{15}$$

To prove this theorem, we need the following lemma. In the sequel, μ_n denotes the empirical measure of $X_1,\ldots,X_n$. Besides, a cube of the form [x-a,x+a] is called a regular cube.

<u>Lemma 3.</u> Let A be a class of regular cubes satisfying the measurability conditions mentioned in paragraph 1 and the condition

$$\operatorname{Sup}_{A\epsilon A} \mu(A) \leq k/n \leq 1/8.$$

Take $k = o(n)$ and

$$k/n \geq \beta \left(\frac{\log n}{n}\right)^{1/(1+2r)}, \tag{16}$$

where $r > 0$ and $\beta > 0$ is any given constant. Then

$$\lim_{n\to\infty} \sup\{(\tfrac{n}{k})^{1+r} \sup_{A\epsilon A} |\mu_n(A) - \mu(A)|\} \leq C_1 \quad \text{a.s.}$$

Notice that A is a V-C class, one can obtain Lemma 3 from Theorem 1 immediately. The proof is omitted.

Proof of Theorem 2. Take $k = o(n)$ and

$$k/n \geq \beta(^{\log n}/n)^{(d+\lambda)/(d+3\lambda)}$$

Put

$$V_n = \theta_1^{-1}(^k/n)^{\lambda/(d+\lambda)}$$

$$q_n = \theta_2 V_n = \theta_1^{-1}\theta_2(^k/n)^{\lambda/(d+\lambda)}$$

$$B_n = \{x: f(x) \geq V_n\}$$

where θ_1, $\theta_2 \varepsilon (0,1)$ will be chosen later.

Let $\mu(x,a)$ and $\mu_n(x,a)$ be the probability measure and empirical measure of $[x-a,x+a]$ respectively. Put $M = \max(\operatorname{Sup}_x f(x),1)$. We have

$$P\{\operatorname{Sup}_{x\varepsilon B_n} |\hat{f}_n(x) - f(x)| > q_n\} \leq I_n + J_n \tag{17}$$

where

$$I_n = P(U_{x\varepsilon B_n}\{\hat{f}_n(x) > f(x) + q_n\}),$$
$$J_n = P(U_{x\varepsilon B_n}\{\hat{f}_n(x) < f(x) - q_n\}). \tag{18}$$

Thus

$$I_n \leq P(U_{x\varepsilon B_n}\{a_n(x) < b_n(x)\}), \tag{19}$$

where

$$2b_n(x) = \{\frac{k}{nf(x)}\,(1+q_n/f(x))^{-1}\}^{1/d}.$$

Fix $x\varepsilon B_n = \{x: f(x) \geq V_n\}$. Take $\theta_2 < 1/8$, then $q_n/f(x) \leq \theta_2 < 1/8$. Noticing $^1/(1+t) < 1 - 7t/8$ for $0 \leq t < 1/8$, we have

$$2b_n(x) \leq \{\frac{k}{nf(x)}\,(1-7q_n/8f(x))\}^{1/d}$$
$$\leq (k/nf(x))^{1/d}.$$

It follows that

$$\mu(x,b_n(x)) = \int_{x-b_n(x)}^{x+b_n(x)} f(t)dt$$

$$\leq (2b_n(x))^d f(x) + C_2(2b_n(x))^{d+\lambda}$$

$$= (2b_n(x))^d f(x)[1 + C_2(2b_n(x))^\lambda/f(x)]$$

$$\leq \frac{k}{n}(1-\frac{7}{8}q_n/f(x))(1+C_2(\frac{k}{nf(x)})^{\lambda/d}/f(x))$$

$$\leq \frac{k}{n}(1-\frac{7}{8}q_n/f(x) + C_2(\frac{k}{nf(x)})^{\lambda/d}/f(x)).$$

Fix θ_2, take θ_1 small enough such that $C_2\theta_1^{(\lambda+d)/d} < \frac{3}{8}\theta_2$, then $C_2(\frac{k}{nf(x)})^{\lambda/d}$ $\leq C_2\theta_1^{\lambda/d}(k/n)^{\lambda/(\lambda+d)} < \frac{3}{8}\theta_1^{-1}\theta_2(k/n)^{\lambda/(d+\lambda)} = \frac{3}{8}q_n$. It follows that

$$\mu(x,b_n(x)) \leq \frac{k}{n}(1-\frac{1}{2}q_n/f(x)) < k/n,$$

and

$$\frac{k}{n} - \mu(x,b_n(x)) \geq kq_n/(2nM).$$

Hence, by (19) and Theorem 1, we have

$$I_n \leq P\{\sup_{x\in B_n}(\mu_n(x,b_n(x)) - \mu(x,b_n(x)) \geq kq_n/(2nM)\}$$

$$\leq C_5 n^\alpha\{\exp(-\frac{n(kq_n/2nM)^2}{91k/n+2kq_n/nM}) + \exp(-k/68)\}$$

where α is a constant depending only on d. In view of (14), we have for large n

$$I_n \leq C_5 n^\alpha\{\exp(-\theta_1^{-1}\theta_2^2 M^{-2}\beta^{1+2\lambda/(d+\lambda)}\log n/400)$$

$$+ \exp(-k/68)\}.$$

Take θ_1 small enough, we have

$$\sum I_n < \infty. \tag{20}$$

In the same way, we can take θ_1 and θ_2 such that

$$\sum J_n < \infty \tag{21}$$

By (17), (18), (20) and (21), we have

$$\sum P\{q_n^{-1} \sup_{x \in B_n} |\hat{f}_n(x) - f(x)| > 1\} < \infty.$$

By Borel-Cantelli's lemma,

$$\lim_{n\to\infty} \sup\{q_n^{-1} \sup_{x \in B_n} |\hat{f}_n(x) - f(x)|\} \leq 1 \text{ a.s.} \tag{22}$$

Fix θ_1, θ_2, and take $2b_n = C_3(k/n)^{1/(d+\lambda)}$. Fix $x \varepsilon B_n^c = \{x: f(x) < V_n\}$. With small C_3 we have

$$\begin{aligned}
\mu(x,b_n) &= \int_{x-b_n}^{x+b_n} f(t)dt \\
&\leq (2b_n)^d f(x) + C_2(2b_n)^{d+\lambda} \\
&\leq \frac{k}{n}[\theta_1^{-1} C_3^d + C_2 C_3{}^{d+\lambda}] < k/2n < k/n.
\end{aligned}$$

Taking $r = \lambda/(d+\lambda)$ in Lemma 3, we can assert with probability one that, for n large enough, the inequality

$$\begin{aligned}
\mu_n(x,b_n) &\leq \mu(x,b_n) + 2C_1(k/n)^{(d+2\lambda)/(d+\lambda)} \\
&< k/2n + 2C_1(k/n)^{(d+2\lambda)/(d+\lambda)} < k/n
\end{aligned}$$

holds uniformly for $x \varepsilon B_n^c$. By definition, for $x \varepsilon B_n^c$,

$$a_n(x) \geq b_n = \frac{1}{2} C_3(k/n)^{1/(d+\lambda)},$$

$$\hat{f}_n(x) \geq C_4(k/n)^{\lambda/(d+\lambda)}$$

It follows that

$$\lim_{n\to\infty} \text{Sup}\{(n/k)^{\lambda/(d+\lambda)} \sup_{x\varepsilon B_n^c} |\hat{f}_n(x) - f(x)|\} \leq C_4 \text{ a.s.} \tag{23}$$

Theorem 2 is proved in view of (22) and (23).

REFERENCES

[1] Chen Xiru and Zhao Lincheng (1984). Almost sure L_1-norm convergence for data-based histogram density estimates. Submitted to J. Multivariate Analysis.

[2] Devroye, L.T. and Wagner, T.J. (1980). The strong uniform consistency of kernel density estimates. Multivariate Analysis - V, North-Holland, 59-77.

[3] Gaenssler, P. and Stute, W. (1979). Empirical processes: a survey of results for independent and identically distributed random variables. Ann. Probability, 7, 193-243.

[4] Vapnik, V.N. and Chervonenkis, A. Ya. (1971). On uniform convergence of the frequencies of events to their probabilities . Theor. Probability Appl., 16, 264-280.

[5] Wenocur, R.S. and Dudley, R.M. (1981). Some special Vapnik-Chervonenkis classes. Discrete Math., 33, 313-318.

[6] Loftsgarden, D.O. and Quesenberry, C.P.(1965). A nonparametric estimate of a multivariate density function. Ann. Math. Statist. 36, 1049-1051.

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFOSR-TR- 85 0978	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): An inequality concerning the deviation between theoretical and empirical distributions	5. TYPE OF REPORT & PERIOD COVERED: Technical - August 1985
	6. PERFORMING ORG. REPORT NUMBER: 85-30
7. AUTHOR(s): Lin-Cheng Zhao	8. CONTRACT OR GRANT NUMBER(s): F49620-85-C-0008
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Center for Multivariate Analysis, 515 Thackeray Hall, University of Pittsburgh, Pittsburgh, PA 15260	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 61102F, 2304/A5
11. CONTROLLING OFFICE NAME AND ADDRESS: Air Force Office of Scientific Research, Department of the Air Force, Bolling Air Force Base, DC 20332	12. REPORT DATE: August 1985
	13. NUMBER OF PAGES: 14
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Empirical distribution, Uniform deviation, Probability inequality, Uniform strong convergence, Nearest neighbor density estimate.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

In this paper the author established an inequality concerning the uniform deviation between theoretical and empirical distributions. An application in strong convergence of nearest neighbor density estimate is also discussed.

[illegible]

DD FORM 1473, 1 JAN 73

END

FILMED

1-86

DTIC